基于维基百科的多种类型文献自动分类研究*

李湘东[1,2] 阮 涛[1] 刘 康[1]
[1](武汉大学信息管理学院 武汉 430072)
[2](武汉大学电子商务研究与发展中心 武汉 430072)

摘要:【目的】通过基于维基百科的特征扩展解决由于不同类型文献而产生的特征不匹配等问题,以提高文本分类效果。【方法】在特征扩展之前,对 TF-IDF 加以改进,提出并使用一种新的特征选择方法 CDF_{max}-IDF 获得候选词集;在使用维基百科进行特征扩展时,通过分别计算直接链接关系、类别关系、间接链接关系三类词语间关系并进行融合得到词语间的语义相关度实现特征扩展;针对扩展得到的特征,提出一种改进的 LDA 概率主题模型 wLDA 模型进行文本建模。【结果】本文提出的方法分别在朴素贝叶斯、KNN 和 SVM 三种分类器上实现分类,其 marco-F1 和 micro-F1 分别提升 1.6%-2.8%和 1.4%-2.7%。【局限】尚未考虑特征词本身及特征词间的相互联系,比如特征词本身的词性、出现在单篇文档中的位置、特征词间的共现关系等因素对特征词权重的影响。【结论】通过多种对比研究证明了使用基于维基百科的特征扩展方法对特征词扩展的有效性,提高了多种类型文献的自动分类效果。

关键词: 多种类型文献 文本分类 特征选择 特征扩展 维基百科
分类号: TP393 G35
DOI: 10.11925/infotech.2096-3467.2017.0702

1 引 言

多种类型文献是指包含了图书、期刊、网页、博客等各种传统和当前流行的社交媒体等形式的文献。从信息管理领域来看,数字图书馆是一种新型图书馆,它既具有传统图书馆在信息整合、组织管理上的优势,又同时可以对来自网络的新兴文本资源(如新闻网页、博客微博等)进行整理收集与分类管理[1];从大数据领域看,其最大特点之一就是数据类型的多样化;除了数值型数据之外,还包含图书、期刊、网页、博客等形式的文本数据。因此,不管是传统意义上的信息资源管理研究,还是当下最前沿的大数据分析,其对象都包括了多种类型文献。

以多种类型文献为研究对象时,一个突出的问题是不同类型文献之间对同一事物或主题使用不同的词汇或特征进行描述、产生语义上的差异,由此导致研究结果的不正确。例如,网页中通常使用的"电脑"可能被大数据分析为是与学术论文中的"计算机"不同的事物或主题。本文以自动分类为手段,通过分类效果的客观比较,找出解决多种类型文献之间语义差异的有效途径。并提出一种基于特征扩展的多种类型文献自动分类方法,通过解决不同类型文献间自动分类时出现的特征不匹配问题,从而消除上述的语义差异,提升多种类型文献自动分类的效果。

2 研究现状及意义

2.1 研究现状

面对高速增长的海量网络信息资源,传统的手工

通讯作者: 李湘东, ORCID: 0000-0001-9031-8482, E-mail: xli_xiao@hotmail.com。
*本文系国家社会科学基金项目"多种类型文本数字资源自动分类研究"(项目编号: 15BTQ066)的研究成果之一。

分类和基于专家系统或知识库的半自动分类方法不能有效地对其进行分类与组织，尤其是面对互联网中多种类型的文本信息，如何有效地对多种类型的信息资源进行有效组织和管理，这对当前的自动文本分类技术提出更高的要求[2]。在自动文本分类的研究领域中，已有相关研究分别以图书书目信息作为训练集、网页新闻文本作为测试集[3]，以期刊论文作为训练集，以期刊论文、网页和图书等三种类型文献作为测试集[4]，以 CiteSeer 中的英文研究论文、学术报告等多种类型文本资源分别作为训练集和测试集[5]，开展多种文献类型的自动分类研究。但是，这些研究对不同类型文献之间可能存在的语义差异未加考虑。

维基百科是目前全球最大的在线协作式百科全书，常常作为第三方知识库引入到研究之中，作为词汇或特征的语义扩展研究中的桥梁被使用。文献[6]以维基百科作为第三方知识库进行特征扩展，使原本只包含较少数量的特征的短文本得以使用语义相近的更多数量的特征加以表达，从而解决短文本分类中存在的特征稀疏等问题，其实验结果证明了维基百科在中文文本语义扩展上的有效性。文献[7]首先使用 LDA (Latent Dirichlet Allocation)模型[8]对英文文本建模并获得特征词，再使用维基百科对所抽象出来的特征词进行语义扩展。文献[9]使用维基百科在来自新闻组和讨论组等两种不同的英文语料之间建立语义关系开展分类，两项研究均通过实验在一定程度上提高了分类效果。因此，维基百科可以有效地用于解决文本中语义扩展或语义差异等问题。

文献[10]将维基百科作为第三方知识库，将其应用到不同类型的中文文献的自动分类之中，通过将来自不同类型文献、语义相近但所使用的词汇不同的特征之间进行扩展和匹配，在一定程度上提高了分类效果。但是，该文献的研究内容有三个方面值得探讨。

(1) 采用传统的 TF-IDF 方法选择特征词；传统的 TF-IDF 是一种基本的特征选择方法、得到广泛的使用；然而，对该方法本身有许多研究并在不断改进[11-13]，值得借鉴。

(2) 使用向量空间模型(Vector Space Model, VSM)进行文本建模；VSM 是一种基本的文本表示模型、得到广泛的应用；然而，VSM 将所有的特征词看作是相互独立的，无法解决同义词、近义词等语义问题；而 LDA 模型可以对文本建模并有效挖掘文本中的语义信息，已经广泛用于包括文本分类在内的各个领域，并取得较好的分类效果[14-17]，可以考虑将其替代 VSM 用于不同类型文献自动分类时的文本表示模型。

(3) 在使用维基百科计算词语相似度时，主要使用了直接链接关系、类别关系、间接链接关系等三类词语间关系，其中，类别关系和间接链接关系分别借鉴了其他论文中的计算方法，相对比较复杂，有进一步简化的可能。

为适应不同领域或学科的多种类型文献之间可能存在的语义差异，本文在三个方面开展了与文献[10]不同的研究，希望提供与文献[10]不尽相同的解决途径。在特征选择方面，对传统 TF-IDF 公式进行改进，引入类间聚集度和类内分散度两个概念，得到改进的 CDF_{max}-IDF 公式进行特征选择；在基于维基百科的相关度计算方法方面，对类别关系和间接链接关系使用简洁的 Jaccard 相似系数公式计算，并对融合公式进一步优化；使用 LDA 模型代替 VSM 进行文本建模，将 LDA 的权重更新公式改进为可以对非整数进行训练，从而得到 wLDA 模型，进行文本表示。

2.2 研究意义

随着网络数字资源的急剧增加，以数字资源为对象的各种研究和应用，例如，大数据分析、数字资源分类等向各个领域快速普及。以数字资源为研究对象时，主要问题在于不同类型间的文献存在一定的语义差异，比如在描述同一事物时，图书期刊等类型的文献偏向于使用事物的学名或规范名称，而新闻网页等类型的文献偏向于使用事物的俗称或常用名，这就导致了在进行文本建模时会出现特征不匹配的问题，而文本建模是大数据分析、数字资源分类的前提和基础，不解决语义差异的问题，意味着文本建模不能反映大数据分析的对象和数字资源分类的对象的实际，必然会影响大数据分析和数字资源分类的结果。本研究通过借助第三方知识库来消除不同文献类型文本间的语义差异，从而提升多种类型文献的自动分类效果，有助于在大数据环境下对多种类型文献进行良好的建模，以实现各种新兴类型和传统类型文献资源的分析、组织与整理工作。本文在这样的背景下进行选题并开展研究，因此具有较高的理论及实用价值。

3　基于维基百科的多种类型文本分类方法

本文大致分为以下步骤进行：对 TF-IDF 加以改进，提出并使用一种新的特征选择方法 CDF_{max}-IDF 获得特征扩展候选词集；基于特征扩展候选词集，对测试集文本使用维基百科中的语义相似度计算进行特征扩展；针对扩展得到的带权特征，提出一种改进的 LDA 概率主题模型 wLDA 模型进行文本建模。其框架如图 1 所示。

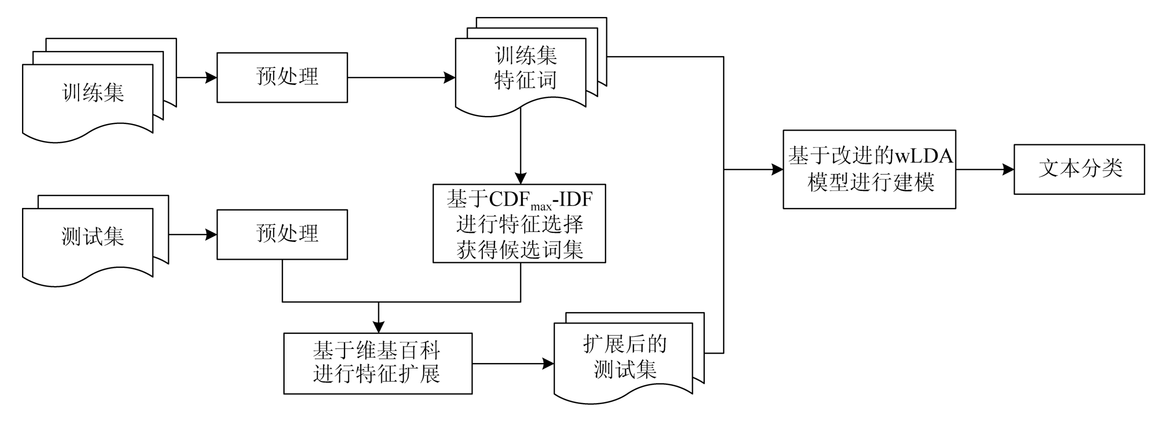


图 1　基于维基百科的多种类型文本分类方法框架

3.1　基于 TF-IDF 特征选择方法的改进

TF-IDF 特征选择方法综合考虑词频和逆文档频率两个因素，认为某特征词在文档中出现的次数越多，且只在很少的文档中出现，说明该特征就越重要[18]。但是传统的 TF-IDF 由于未考虑特征的分布情况而未能选择出更具代表性特征。文献[11]针对传统 TF-IDF 的不足，提出类间分散度和类内分散度两个概念并将其与传统 TF-IDF 相结合，对基于 TF-IDF 的特征加权算法进行改进；文献[12]引入类间集中度和类内分散度两个重要概念，并对文献[11]中的公式进行重新定义，主要区别在于计算时仅使用特征项的文档频率作为唯一参数，减少了计算的复杂程度，且通过实验验证其与文献[11]方法同样可以选择出更具类别区分能力的特征。最后，再结合频度因素提出一种改进的基于 TF-IDF 的特征加权算法。该算法认为如果某一个特征项的频度越高、在类别间分布越聚集且在类别内分布越分散，那么该特征对文本类别的区分作用就越大，即分辨度越强。文献[13]将文献[12]的方法应用到 LDA 主题模型中，进一步验证了文献[12]提出的改进的 TF-IDF 特征加权算法的有效性。但是，文献[12-13]在计算类间聚集度时，由于使用的是文档频率而未能反映出特征频度。因此，本文在保留文献[12-13]的类间聚集度和类内分散度两个概念的基础上，使用基于特征词频的类间聚集度代替文献[12-13]中基于文档频率的类间聚集度。

类间聚集度和类内分散度主要从特征词在类别间以及类别内分布情况的角度进行特征权重的考量，克服了传统 TF-IDF 在计算特征权重时没有综合考虑类间、类内特征分布所存在的不足。而基于特征词频的类间聚集度不仅可以反映类别间的分布情况，还可以反映出频度信息。所以本文在最终改进的 TF-IDF 中去掉了特征词频 $tf_i(t)$，这样有助于降低算法的计算复杂度。通过引入基于特征词频的类间聚集度和类内分散度两个参数，本文提出一种改进的 CDF_{max}-IDF 特征选择算法，其计算公式如下：

$$CDF_{max}\text{-}IDF(t)=\max_{i=1}^{k}(IC_i(t)\cdot ID_i(t))\cdot\log\left(\frac{|D|}{\sum_{i=1}^{k}df_i(t)+\varepsilon}\right) \tag{1}$$

其中，k 表示总类别数，$|D|$表示文本集中文档总数，$df_i(t)$表示在第 i 个类别中特征词 t 的文档频数，ε 为平滑因子，$IC_i(t)$表示类别 C_i 下特征词 t 的基于特征词频的类间聚集度，$ID_i(t)$表示类别 C_i 下特征词 t 的类内分散度，对其乘积结果按类别取最大值。

3.2　基于维基百科的特征扩展方法

(1) 基于维基百科的语义相关度计算

①直接链接关系

维基百科中的任意一个概念的解释页面中存在大量其他概念的引用，引用概念和被引用概念之间往往存在极强

的相关性，因此概念间是否存在链接关系通常被作为衡量概念间相关度的一项重要指标[19]。如果概念 t_1 的解释页面中含有概念 t_2 的链接，则称概念 t_1 是概念 t_2 的链入链接或概念 t_2 是概念 t_1 的链出链接。此时，若概念 t_2 也是 t_1 的链入链接，则称概念 t_1 与概念 t_2 之间存在双链接关系，反之则称为单链接关系。

②链接相关度

链接相关度是通过衡量维基百科中任意两项概念拥有共同链入、链出概念的数量及其相互间的覆盖程度来确定的概念间的相关度。维基百科中的每一项概念都拥有两个相关的链接集合——链出概念集和链入概念集，分别是由该概念的链出概念和链入概念构成。利用 Jaccard 相似系数可以很容易地衡量两个集合之间的相似性[20]，其计算公式如下：

$$JS(A,B)=\frac{|A\cap B|}{|A\cup B|} \tag{2}$$

其中，A、B 分别代表两个集合。$A\cap B$ 表示 A、B 两集合的交集；$A\cup B$ 表示 A、B 两集合的并集。

本文基于 Jaccard 相似系数计算两个维基概念的链接相关度。首先分别从维基百科中查找出概念 t 的链入概念集和链出概念集：

$$\begin{aligned}&inlinks(t)=\{inlink_{t1},inlink_{t2},inlink_{t3},\cdots\}\\&outlinks(t)=\{outlink_{t1},outlink_{t2},outlink_{t3},\cdots\}\end{aligned} \tag{3}$$

通常认为一个概念的链入概念和链出概念都是该概念的相关概念，即链入概念集和链出概念集都具有概念表征能力，因此基于 Jaccard 相似系数得到的概念链接相关度计算公式如下：

$$sim_{link}(t_i,t_j)=\alpha\cdot\frac{|inlinks(t_i)\cap inlinks(t_j)|}{|inlinks(t_i)\cup inlinks(t_j)|}+\beta\cdot\frac{|outlinks(t_i)\cap outlinks(t_j)|}{|outlinks(t_i)\cup outlinks(t_j)|} \tag{4}$$

其中，α和β是权重参数，满足$\alpha+\beta=1$，本文利用其计算概念间相关度时，认为链入概念集和链出概念集拥有同样的表征，故选取$\alpha=\beta=0.5$。

③类目相关度计算

由于维基百科中的任意一项概念都至少属于一个类目下，换句话说，每一个维基概念都会拥有一个属于它的所属类目集：

$$categories(t)=\{category_{t1},category_{t2},category_{t3},\cdots\} \tag{5}$$

因此，同样可以利用 Jaccard 相似系数计算方法实现两个概念间类目相关度的计算，但值得注意的是，在维基百科中的类目存在包含与被包含的关系，且层级越高的类目往往拥有更多的从属概念，所以不能直接使用 Jaccard 相似系数表征两个概念间的类目相关度。本文在计算类目相关度时，对 Jaccard 相似系数进行加权改进，以此平衡不同层级类目在计算相关度时的权重问题，其具体计算公式如下：

$$\begin{cases}sim_{cate}(t_i,t_j)=\dfrac{\sum_{c}^{categories(t_i)\cap categories(t_j)}weight(c)}{\sum_{c}^{categories(t_i)\cup categories(t_j)}weight(c)}\\[2ex] weight(c)=\dfrac{1}{n_c}\end{cases} \tag{6}$$

其中，n_c 代表类目 c 的从属概念数量，即用类目的从属概念数量的倒数替代原始公式中的类目本身，从而对 Jaccard 相似度计算公式进行加权，这么做可以让共同从属于较低层级类目的两概念之间获得更高的相关度计算结果。

④概念相关度计算

概念相关度计算是指对任意两个存在于维基百科中的概念，通过定义一个合理的计算公式以准确度量两个概念间的语义相关度，并且要求该相关度计算结果高的概念间需要能够从很大程度上说明这两个概念具有非常紧密的语义联系或通常在描述同一特定领域时共现[21]。综合考虑上述概念之间的三种关系的相关度计算公式，本文提出一种新的概念相关度方法：

$$\begin{aligned}&sim(t_1,t_2)=\max(\delta(\alpha\cdot sim_{link}(t_1,t_2)+\beta\cdot sim_{cate}(t_1,t_2)),1)\\&\delta(x_{t_1,t_2})=\begin{cases}\eta_1 x & t_1,t_2\text{存在双链接关系}\\ \eta_2 x & t_1,t_2\text{存在单链接关系}\\ x & t_1,t_2\text{不存在链接关系}\end{cases}\end{aligned} \tag{7}$$

其中，α、β为可调参数，且$\alpha+\beta=1$，用来调节链接相关度与类目相关度的参考权重。$\delta(x)$是关于概念 t_1、t_2 直接链接关系的示性函数，系数 η_1、η_2 取值大于 1，代表对拥有直接链接关系的两项维基概念的相关度计算结果进行不同程度的加权，这样能尽可能保证拥有直接链接关系的概念能够被准确赋予更高的相关度。

(2) 基于语义相关度的特征扩展方法

基于语义相关度的特征扩展方法主要步骤如下：

①对整个文本集中的文本进行分词、词性过滤、停用词过滤等操作；

②对训练集文本进行特征选择，得到特征扩展候选词集；

③对测试集中的每一篇文档，利用提出的基于维基百科的相关度计算方法，依次计算其与特征扩展候选词集中各特征词的相关度，完成特征扩展。

3.3 基于 wLDA 模型的文本分类方法

(1) 标准 LDA 模型

LDA 模型是 Blei 在 PLSI 模型的基础上，引入贝叶斯思想后提出的一种全新的概率生成模型。LDA 模型在文本生成模型中引入了多项分布的共轭先验分布

——狄里克雷(Dirichlet)分布，从而构建了一个从词到主题，再从主题到文档的三层结构概率文本表示模型。LDA 主题模型的概率模型图[22]如图 2 所示。

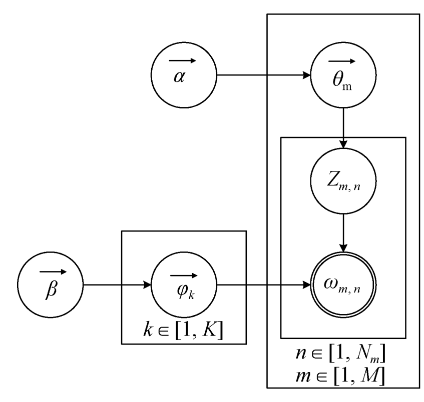


图 2 LDA 主题模型的概率模型图

其中，M 表示文档总数，N_m 表示第 m 篇文档中存在 N_m 个特征词，$\overrightarrow{\theta_m}$ 代表文档–主题的概率分布，$\overrightarrow{\varphi_k}$ 代表了主题–词的概率分布，K 表示主题总个数，$\vec{\alpha}$ 和 $\vec{\beta}$ 分别是两个分布的超参数，$Z_{m,n}$ 是由分布 $\overrightarrow{\theta_m}$ 生成的第 m 篇文档中第 n 个词即 $\omega_{m,n}$ 的所属主题。

(2) LDA 主题模型的求解及评价

①模型求解

LDA 模型通常采用吉布斯采样方式估计特征词 ω 和主题 z 的后验分布。吉布斯采样的计算公式[8]为:

$$p(z_i = k \mid \vec{z}_{\neg i}, \vec{w}) \propto \frac{n_{m,\neg i}^{(k)} + \alpha^{(k)}}{\sum_{j=1}^{K}(n_{m,\neg i}^{(j)} + \alpha^{(j)})} \cdot \frac{n_{k,\neg i}^{(t)} + \beta^{(t)}}{\sum_{j=1}^{V}(n_{k,\neg i}^{(j)} + \beta^{(j)})} \quad (8)$$

其中，z_i 表示第 i 个特征词对应的主题变量，$\vec{z}_{\neg i}$ 表示 $k \neq i$ 时特征词编号的主题分布，$n_{m,\neg i}^{(k)}, n_{k,\neg i}^{(t)}$ 分别代表第 m 篇文档中主题 k 的频数以及主题 k 中特征词 t 的频数。

根据狄里克雷分布的参数估计公式可得:

$$\hat{\theta}_{mk} = \frac{n_{m,\neg i}^{(k)} + \alpha^{(k)}}{\sum_{j=1}^{K}(n_{m,\neg i}^{(j)} + \alpha^{(j)})}$$
$$\hat{\varphi}_{kt} = \frac{n_{k,\neg i}^{(t)} + \beta^{(t)}}{\sum_{j=1}^{V}(n_{k,\neg i}^{(j)} + \beta^{(j)})} \quad (9)$$

$\hat{\theta}_{mk}, \hat{\varphi}_{kt}$ 分别代表第 m 篇文档中选取主题 k 的概率估计以及主题 k 中选取特征词 t 的概率估计。

②模型评价

LDA 主题模型本质上是一种文本聚类算法，所以在利用 LDA 进行文本建模的时候，需要事先指定主题数，常用的指标有主题相似度[23]、困惑度[24]等，本研究将选择困惑度作为确定最优主题数的方法。随着主题数的逐渐增大，困惑度会一直减小，但最终其值会趋于平稳。因此通过综合考虑困惑度和主题数的比率，可以在合理范围内选取最合适的主题数，得到最优的文本概率模型。

(3) 改进的 LDA 模型

标准 LDA 主题模型在通过吉布斯采样训练模型参数的时候，是利用特征词的频数对参数进行迭代更新，这样会使文档的主题明显倾向于高频词的主题分布，从而影响分类效果。另一方面，根据上文所述的特征扩展方法得到的扩展词以[0,1]的相关度被扩展到文本集中、形成带有权值的扩展特征，这样的参数更新方式并不能对这些带有权值的扩展特征进行吉布斯采样，因此有必要对标准 LDA 模型的求解方式进行改进，改进后的吉布斯采样公式为:

$$p(z_i = k \mid \vec{z}_{\neg i}, \vec{w}) \propto \frac{weight(n_{m,\neg i}^{(k)}) + \alpha^{(k)}}{\sum_{j=1}^{K}(weight(n_{m,\neg i}^{(j)}) + \alpha^{(j)})} \cdot \frac{weight(n_{k,\neg i}^{(t)}) + \beta^{(t)}}{\sum_{j=1}^{V}(weight(n_{k,\neg i}^{(j)}) + \beta^{(j)})} \quad (10)$$

利用改进的 weight-LDA(简称 wLDA)模型，可以对特征扩展后的文本进行主题建模，从而使用分类算法进行文本分类。

(4) 基于 wLDA 模型的分类流程

结合前文的特征扩展方法，本文提出结合特征扩展和改进的 wLDA 主题模型的文本分类算法，其具体流程如下:

①对训练集和测试集的文本进行分词、词性过滤、停用词过滤等操作;

②对训练集文本进行初步统计，通过本文提出的 CDF_{max}-IDF 特征选择算法筛选出特征扩展候选词集，并通过维基词典进行特征过滤，即选取在维基百科中能找到对应概念的特征词;

③针对测试集中的每一篇文档，利用基于维基百科的相关度计算方法依次计算文档中的每一项特征词和特征扩展候选词集中每一项特征词的相关度，将相关度大于某一阈值的特征词扩展到测试集文本中，形成包含带有权值的扩展特征的新测试集文本;

④使用常见的几种分类算法对训练集建模，对测试集进行分类并评价分类效果。

4 实验及结果分析

4.1 实验准备

本实验利用期刊论文为主的文献信息构成训练语料，对以新闻网页为主要类型的文献信息进行分类实验，实现多种文献类型的文本自动分类。为使实验过程与结果满足公开原则与可复现性，本文实验所用语料全部取自复旦大学中文语料库[25]和搜狗互联网语料库[26]，其中复旦语料库以期刊文本和学术论文为主，搜狗语料库主要由新闻网页文本组成。本实验选取复旦语料库与搜狗语料库中共有的经济、体育、环境三个类别语料各 800 篇，其中复旦语料库中各抽取 500 篇组成训练语料(共 1 500 篇)，搜狗语料库中各抽取 300 篇组成测试语料(共 900 篇)，为避免随机干扰项对实验结果造成影响，在公开语料集上利用多次随机抽取的文本集进行实验，对实验数据取平均值作为最终的实验结果。

4.2 实验设计

(1) LDA 主题模型中最优主题数的确定:其基本过程为对训练集在不同主题数下进行多次 LDA 主题建模，并分别计算其困惑度，再根据困惑度的变化趋势确定最合理的主题个数。

(2) 扩展特征语义相关度计算结果的考量。本文解决多种文献类型文本分类的主要思路是通过特征扩展来消除训练集与测试集间的语义差异，而特征扩展的核心则是计算特征间的语义相关度，但对语义相关度计算结果的考量并没有一种权威的测度方式，因此将随机抽取几组特征的语义相关度计算结果，通过人工审查的方式对计算结果是否合理做出合理的定性判断。

(3) 本文分类方法分类结果的对比验证。为了验证本文提出的基于特征选择的多种文献类型文本自动分类方法的有效性，通过在多种经典的自动分类算法上进行分类实验，分别对是否使用本文提出的分类方法的分类效果进行比较分析，从而论证本文提出的改进方法的可行性和有效性。

4.3 实验结果及分析

(1) 最优主题数的确定

本文使用困惑度来确定 LDA 模型的最优主题数，使用吉布斯采样法求解 LDA 模型的参数，其中超参数 α、β 根据经验分别设置为 $50/t$(t 为主题数)和 0.01，对训练集迭代次数为 1 000 次。本实验中对主题数进行从 10 到 150 的预设(梯度为 10)分别计算该主题数下的困惑度，绘制成趋势图如图 3 所示。

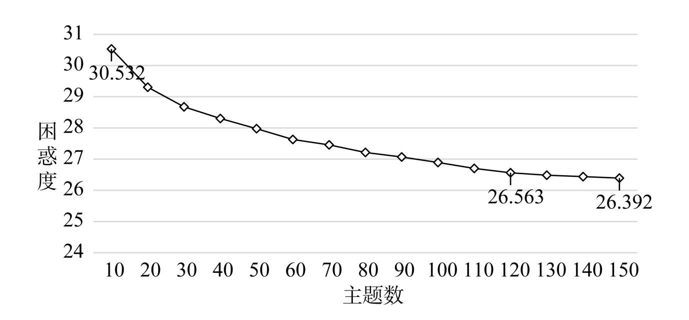


图 3 训练集在 LDA 模型下的主题数变化趋势

由图 3 可知，困惑度随着主题数逐渐增大而下降，且下降趋势逐渐平稳，本文选取困惑度下降趋势明显放缓的第一个拐点作为本实验中 LDA 模型的最优主题数，因此选取 120 作为后续 LDA 建模的最优主题数。

(2) 扩展特征语义相关度计算结果

特征扩展是本文解决多种文献类型文本分类方法的核心环节，而语义相关度计算结果的准确与否是决定特征扩展最终效果的直接影响因素。在进行特征扩展前，先分别使用 TF-IDF 方法和本文提出的 CDF_{max}-IDF 特征选择方法从训练集中提取特征扩展候选词集。表 1 为一次实验中使用 TF-IDF 特征选择方法提取的候选词集(取前 45 个特征词)。

表 1 基于 TF-IDF 方法的特征扩展候选词集

关键词
经济、体育、企业、发展、市场、浓度、社会、政府、产业、改革、增长、投资、我国、土壤、国有、消费、制度、地区、吸附、技术、图、结构、政策、中国、工业、降解、专业、农村、资本、水、管理、菌、国家、农业、知识、污泥、生产、要、研究、产品、教育、环境、体制、氧、人、……

使用传统 TF-IDF 方法进行特征选择得到的候选词集中存在许多诸如“中国”、“要”、“人”这类相对高频但却不具有类别区分能力的特征词，如果将这类特征词作为候选词会在特征扩展的同时引入大量的“噪声”而影响分类结果。表 2 是则使用本文提出的 CDF_{max}-IDF 特征选择方法提取的候选词集(每个类各取前 15 个特征词)，通过对比可以说明本文提出的特征选择方法的有效性。

表 2　基于 CDF_{max}-IDF 的特征扩展候选词集

类别	关键词
经济	资本、经济增长、企业、经济发展、市场、政策、金融、价格、投资、增长、资金、国民经济、利益、劳动力、市场经济、……
体育	比赛、队、体育、运动员、冠军、选手、成绩、队员、女子、速率、决赛、训练、胜、力量、中国队、……
环境	环境科学、浓度、中国环境、scientiae、水、污染、污染物、化学、温度、试验、生物、离子、含量、pollution、监测、……

表 2 对筛选得到的类别关键词进行了类别区分，便于观察特征词是 CDF_{max}-IDF 计算结果取最大值时的所属类别以及其类别归属本身的合理性。在实际实验过程中利用关键词集进行特征扩展时是不区分其所属类别的，而是将所有关键词汇总在一起不作区分地计算语义相关度并与阈值进行比较确定是否被扩展。表 3 是一次实验中对待分类文本的几个特征词进行特征扩展时得到的语义相关度计算结果。

表 3　语义相似度计算结果

特征词	扩展特征词及语义相关度
市场	交易:0.102　金融市场:0.211　劳动力市场:0.212　批发:0.224
股东	股票市场:0.146
净利润	资金:0.111　增长率:0.136　市场化:0.108　负债:0.172
女排	排球:1.000
王宝泉	袁伟民:0.115
亚军	冠军:0.709　金牌:0.106　银牌:0.274
环境监测	污染:0.346　污染物:0.100　富营养化:0.148
凝固	蒸发:0.288
污水处理	水质:0.173　水污染:0.357　生活污水:0.112

表 3 从三个类中各列举了三个特征词扩展结果作为代表，不难看出扩展得到的特征词确实与原特征词具备很强的语义关联，如“市场”与“交易”、“金融市场”、“劳动力市场”等。与此同时也可以充分反映出不同扩展特征词与原特征词间的关联强弱差异，比如“亚军”与“冠军”、“金牌”、“银牌”间的语义相关度为“冠军”>“银牌”>“金牌”，这基本符合一般认知。因此，利用维基百科进行的语义相关度计算方法，不仅能对互为同义词、近义词的特征词进行扩展，还能利用维基百科中蕴含的领域知识，将类似“王宝泉”与“袁伟民”这样只有具备相当领域背景知识才能正确识别和处理的相关特征词进行扩展。

通过以上分析可以证明本文提出的基于维基百科的特征扩展方法确实能够对待分类文本进行恰当的特征扩展处理，接下来再通过分析对比最终的分类实验结果论述特征扩展后得到的新的待分类文本可以更容易地被正确自动分类。

(3) 本研究分类方法的对比验证

为了验证本文提出的基于特征选择的多种文献类型文本自动分类方法的有效性，通过在三种经典的自动分类算法(K 最近邻算法、朴素贝叶斯算法、支持向量机算法)上进行分类实验，分别比较是否使用本文提出的分类方法的效果，如图 4 和图 5 所示。

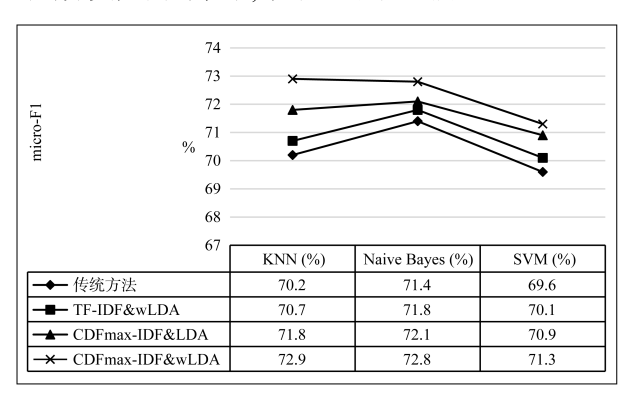


图 4　三种分类算法上的 micro-F1 分类结果对比

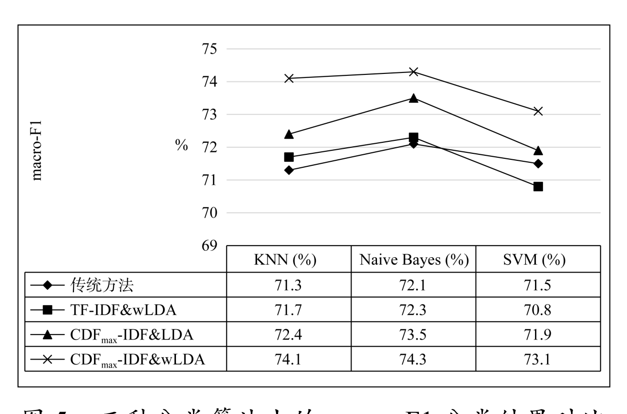


图 5　三种分类算法上的 macro-F1 分类结果对比

根据图 4、图 5 展示的实验结果，分别使用 macro-F1 和 micro-F1 两种评价指标对本研究中涉及的对比实验进行评价。在三种经典分类算法上分别比较特征扩展前后的分类结果，其中朴素贝叶斯算法的平均表现最好，macro-F1 和 micro-F1 在使用本文改进的特征扩展方法后准确率分别达到了 74.3%和 72.8%，

较未进行扩展的结果分别提升了 2.2%和 1.4%；而在使用了本文提出的改进方法后使用 KNN 算法进行分类的 marco-F1 和 micro-F1 分别提升了 2.8%和 2.7%，是三种分类算法中提升最为明显的；在使用 SVM 算法的分类结果上，本文提出的特征扩展方法 marco-F1 和 micro-F1 分别提升了 1.6%和 1.7%。同时为证明本文提出改进的特征选择方法和wLDA主题模型的有效性，还分别设置了 TF-IDF&wLDA 方法及 CDF_{max}-IDF&LDA 方法的对照组进行实验，结果表明本文提出的新的文本分类方法均优于另外两种只进行其中一种改进的分类方法。综上，本文提出的基于特征扩展的多种类型文献自动分类方法是可行且有效的。

5 结 语

本文提出一种基于特征扩展的多种类型文献分类方法，通过利用维基百科作为第三方知识库消除不同文献类型文本间的语义差异，由此提高多种类型文献混合自动分类的分类效果。针对传统 TF-IDF 的不足，对 TF-IDF 加以改进，提出并使用一种新的特征选择方法 CDF_{max}-IDF 获得特征扩展候选词集；在使用维基百科进行特征扩展时，通过分别计算直接链接关系、类别关系、间接链接关系三类词语间关系并进行融合得到词语间的语义相关度进行特征扩展；针对扩展得到的带有权值的特征，提出一种改进的LDA概率主题模型 wLDA 模型进行文本建模，使特征词被赋予了不同权重，提高了 LDA 模型本身的精度和准确性。

本文通过实验论证了提出的基于维基百科的特征扩展方法能在一定程度上提高多种类型文献自动分类的分类效果，但本方法实际上还存在诸多局限性，比如改进后的 CDF_{max}-IDF 特征选择方法虽然在继承了 TF-IDF 优点的基础上又综合考虑了特征词在各个类别中出现的分布和在整个文档集中出现的文档频次，但缺乏对特征词本身及特征词间的相互联系的充分考虑，比如特征词本身的词性、出现在单篇文档中的位置、特征词间的共现关系等，这些都能对特征词的权重起到一定的度量作用，可予以适当考虑。此外，将本方法应用于英文语料也是一个需要检验的课题。

参考文献：

[1] 和艳会，李和娟，关琼，等. 浅谈网络图书馆、数字图书馆、虚拟图书馆的概念[J]. 农业图书情报学刊, 2006, 18(9): 120-123. (He Yanhui, Li Hejuan, Guan Qiong, et al. Discussion on Concepts of Network Library, Digital Library and Virtual Library [J]. Journal of Library and Information Sciences in Agriculture, 2006, 18(9): 120-123.)

[2] 李湘东，胡逸泉，巴志超，等. 数字图书馆多种类型文献混合自动分类研究[J]. 图书馆杂志，2014，33(11)：42-48. (Li Xiangdong, Hu Yiquan, Ba Zhichao, et al. The Study of Mixed Automatic Categorization on Digital Library Collections [J]. Library Journal, 2014, 33(11): 42-48.)

[3] Pong J Y-H, Kwok R C-W, Lau R Y-K, et al. A Comparative Study of Two Automatic Document Classification Methods in a Library Setting[J]. Journal of Information Science, 2008, 34(2): 213-230.

[4] 薛春香，夏祖奇，侯汉清. 基于语料和基于标引经验的自动分类模式比较[J]. 南京农业大学学报：社会科学版，2005, 5(4): 85-91. (Xue Chunxiang, Xia Zuqi, Hou Hanqing. A Comparison of Automatic Classification Between Corpus-based Model and Experiences-based Model [J]. Journal of Nanjing Agricultural University: Social Sciences Edition, 2005, 5(4): 85-91.)

[5] Joorabchi A, Mahdi A E. An Unsupervised Approach to Automatic Classification of Scientific Literature Utilizing Bibliographic Metadata[J]. Journal of Information Science, 2011, 37(5): 499-514.

[6] 范云杰，刘怀亮. 基于维基百科的中文短文本分类研究[J]. 现代图书情报技术，2012(3)：47-52. (Fan Yunjie, Liu Huailiang. Research on Chinese Short Text Classification Based on Wikipedia [J]. New Technology of Library and Information Service, 2012(3): 47-52.)

[7] Guo N, He Y, Yan C G, et al. Multi-level Topical Text Categorization with Wikipedia [C]// Proceedings of International Conference on Utility and Cloud Computing. ACM, 2016: 343-352.

[8] Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation[J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[9] Peter, Maxwell. Co-Clustering Based Classification Algorithm with Latent Semantic Relationship for Cross-Domain Text Classification Through Wikipedia[J]. Bonfring International Journal of Data Mining, 2017, 7(2): 1-5.

[10] 李湘东，刘康，高凡. 维基百科在多种类型数字文本资源自动分类中的应用[J]. 情报科学，2017，35(2)：75-79. (Li Xiangdong, Liu Kang, Gao Fan. Application of Wikipedia to Automatic Categorization with Multiple Types of Digital Text Resources[J]. Information Science, 2017, 35(2): 75-79.)

[11] 徐凤亚, 罗振声. 文本自动分类中特征权重算法的改进研究[J]. 计算机工程与应用, 2005, 41(1): 181-184. (Xu Fengya, Luo Zhensheng. An Improved Approach to Term Weighting in Automated Text Classification[J]. Computer Engineering and Applications, 2005, 41(1): 181-184.)
[12] 蒋健. 文本分类中特征提取和特征加权方法研究[D]. 重庆: 重庆大学, 2010. (Jiang Jian. Research on Feature Extraction and Feature Weighting in Text Categorization[D]. Chongqing: Chongqing University, 2010.)
[13] 李湘东, 丁丛, 高凡. 基于复合加权 LDA 模型的书目信息分类方法研究[J]. 情报学报, 2017, 36(4): 352-360. (Li Xiangdong, Ding Cong, Gao Fan. The Research of Bibliographic Information Classification Method Based on the Composite Weighted LDA Model[J]. Journal of the China Society for Scientific andTechnical Information, 2017, 36(4): 352-360.)
[14] 李锋刚, 梁钰, GAO Xiaozhi, 等. 基于 LDA-wSVM 模型的文本分类研究[J]. 计算机应用研究, 2015, 32(1): 21-25. (Li Fenggang, Liang Yu, GAO Xiaozhi, et al. Research on Text Categorization Based on LDA-wSVM Model[J]. Application Research of Computers, 2015, 32(1): 21-25.)
[15] Li X, Ouyang J, Zhou X, et al. Supervised Labeled Latent Dirichlet Allocation for Document Categorization[J]. Applied Intelligence, 2015, 42(3): 581-593.
[16] 史庆伟, 从世源. 基于mRMR和LDA主题模型的文本分类研究[J]. 计算机工程与应用, 2016, 52(5): 127-133. (Shi Qingwei, Cong Shiyuan. Research on Text Categorization Based on mRMR and LDA [J]. Computer Engineering and Applications, 2016, 52(5): 127-133.)
[17] Lin W, Pang X, Wan B, et al. MR-LDA: An Efficient Topic Model for Classification of Short Text in Big Social Data[J]. International Journal of Grid & High Performance Computing, 2016, 8(4): 100-113.
[18] 孙建军. 信息检索技术[M]. 北京: 科学出版社, 2004: 169-170. (Sun Jianjun. Information Retrieval Technology [M]. Beijing: Science Press, 2004: 169-170.)
[19] 王兰成, 刘晓亮. 维基百科知网的构建研究与应用进展[J]. 情报资料工作, 2012(5): 56-60. (Wang Lancheng, Liu Xiaoliang. Construction Research and Application Progress of Wikipedia Knowledge Network [J]. Information and Documentation Services, 2012(5): 56-60.)
[20] 卢盛祺, 管连, 金敏, 等. LDA 模型在网络视频推荐中的应用[J]. 微型机与应用, 2016, 35(11): 74-79. (Lu Shengqi, Guan Lian, Jin Min, et al. The Application of LDA in Online Video Recommendation [J]. Microcomputer and Its Applications, 2016, 35(11): 74-79.)
[21] 周琨峰. 基于中文维基百科的概念相关词群研究[D]. 武汉: 华中师范大学, 2012. (Zhou Kunfeng. Research on the Concept-related Phrases Based on Chinese Wikipedia [D]. Wuhan: Huazhong Normal University, 2012.)
[22] Wei X, Croft W B. LDA-based Document Models for Ad-Hoc Retrieval[C]//Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2006: 178-185.
[23] 王振振, 何明, 杜永萍. 基于 LDA 主题模型的文本相似度计算[J]. 计算机科学, 2013, 40(12): 229-232. (Wang Zhenzhen, He Ming, Du Yongping. Text Similarity Computing Based on Topic Model LDA [J]. Computer Science, 2013, 40(12): 229-232.)
[24] Cao J, Xia T, Li J, et al. A Density-based Method for Adaptive LDA Model Selection[J]. Neuro Computing, 2009, 72(7): 1775-1781.
[25] 复旦大学中文语料库[DB/OL]. [2017-03-01]. http://www.datatang.com/data/43318. (Fudan-Classification-Corpus [DB/OL]. [2017-03-01]. http://www.datatang.com/data/43318.)
[26] 搜狗互联网语料库[DB/OL]. [2017-03-01]. http://www.sogou.com/labs/. (SogouT [DB/OL]. [2017-03-01]. http://www.sogou.com/labs/.)

作者贡献声明:

李湘东: 提出研究思路, 设计研究方案, 论文最终版修订;
阮涛: 采集数据, 结果分析, 论文修订;
刘康: 论文起草, 进行实验。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据见期刊网络版 http://www.infotech.ac.cn。
[1] 李湘东, 阮涛, 刘康. data.zip. 训练集和测试集文件.
[2] 李湘东, 阮涛, 刘康. collection.zip. 实验中间文件.
[3] 李湘东, 阮涛, 刘康. result.zip. 实验结果文件.

收稿日期: 2017-07-17
收修改稿日期: 2017-08-17

Automatic Classification of Documents from Wikipedia

Li Xiangdong[1, 2] Ruan Tao[1] Liu Kang[1]
[1](School of Information Management, Wuhan University, Wuhan 430072, China)
[2](Center for Electronic Commerce Research and Development, Wuhan University, Wuhan 430072, China)

Abstract: [**Objective**] This paper aims to improve the performance of text classification systems with the help of Wikipedia's feature expansion function. [**Methods**] First, we established the CDF_{max}-IDF method based on the modified TF-IDF, which helped retrieve the candidate word list. Then, we used the Wikipedia to extend the document features and calculated the relationship among direct links, categories and indirect links, which decided the semantic relevance of the words. Finally, we proposed an improved LDA model, the wLDA, for the extended feature and text modeling. [**Results**] The proposed method improved the value of marco-F1 and micro-F1 on Naive Bayes, KNN and SVM classifiers by 1.6%-2.8% and 1.4%-2.7%. [**Limitations**] We did not include the properties of the words and relationship among them. [**Conclusions**] The feature expansion method based on the Wikipedia improves the effectiveness of automatic document classification methods.

Keywords: Various Types of Documents Text Classification Feature Selection Feature Expansion Wikipedia

直布罗陀国家档案馆与 Preservica 合作进行数字资源长期保存

近日，直布罗陀国家档案馆(Gibraltar National Archives, GNA)宣布与 Preservica 合作，保存并保护该国广泛收集的历史数字资源。通过使用 Preservica 的数字保存软件和协同工作方法，GNA 的历史记录可以安全地存储几十年供未来多代后人使用。

2014 年 GNA 重新命名后，创建了一个网站，并升级成为 21 世纪流行的存储库，储存内容包含物理的、数字的和原生数字的记录。下一步是进行长期保存，需要选择一个系统，对归档的、有长期保存价值的、从纸质记录转换而成的数字记录进行永久的保存和保护。这些数字档案包括数千个二次世界大战撤离记录，1502 年直布罗陀的图像，几幅世界名画和重要制图。

直布罗陀首席副部长 Joseph Garcia 博士说：“我们的国家档案馆里的档案是至关重要的，它记录了直布罗陀作为一个国家的历史进程，那些地图、文件、照片和其他文物，是我们国家历史的一个形象的表述，这些档案将永远把这段历史带入到直布罗陀人的生命之中。”

英国国家档案馆为 GNA 推荐了 Preservica 的数字保存和访问软件。GNA 负责人 Anthony Pitaluga 在英国国家档案馆参加了一些数字保存研讨会，阅读了相关文件，并最终选定了 Preservica 系统。

GNA 选择了在 AWS 上托管的 Preservica 云版本，以满足对直布罗陀丰富的数字存档的摄取、处理、安全存储、管理和访问等所有要求。此外，使用云中托管的保护和访问系统意味着档案不需要购买本地服务器和存储，从而成为一种非常具有成本效益且价格合理的选择。使用 Preservica, GNA 建立了一个可搜索的数据库，提取并导入了 20 多万条记录。

(编译自: https://preservica.com/resources/press-releases/gibraltar-national-archives-chooses-preservica-to-safeguard-its-rich-heritage-1)

(本刊讯)